Analise exploratória

A analise será feita no datasete [bfd\_2019](%22https://raw.githubusercontent.com/eogasawara/datamining/main/data-work/bfd_2019.rdata%22).

O objetivo é analisar e falar sobre o que tem na base assim como detalhes importantes e descobertas, assim gerando perguntas sobre o tema

A base consiste em 983396 contendo 48 colunas, a documentação sobre os atributos estão em [documentação](https://osf.io/8eh3p/)

carregando a basse e as lib que serão utilizadas

dados <- get(load(url("https://raw.githubusercontent.com/eogasawara/datamining/main/data-work/bfd\_2019.rdata")))

obtendo algumas informações sobre a base

dim(dados)

## [1] 983396 48

str(dados)

## 'data.frame': 983396 obs. of 48 variables:  
## $ arrival : chr "CYUL" "CYUL" "CYUL" "CYUL" ...  
## $ depart : chr "SBGR" "SBGR" "SBGR" "SBGR" ...  
## $ route : chr "SBGR-CYUL" "SBGR-CYUL" "SBGR-CYUL" "SBGR-CYUL" ...  
## $ company : chr "ACA" "ACA" "ACA" "ACA" ...  
## $ flight : chr "0097" "0097" "0097" "0097" ...  
## $ di : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ type : chr "I" "I" "I" "I" ...  
## $ depart\_day\_period : Ord.factor w/ 7 levels "Night"<"Early Morning"<..: 2 2 2 2 2 2 2 2 2 7 ...  
## $ arrival\_day\_period : Ord.factor w/ 7 levels "Night"<"Early Morning"<..: 6 6 6 6 6 6 6 6 6 2 ...  
## $ expected\_depart : POSIXlt, format: "2019-12-12 08:40:00" "2019-12-14 08:40:00" ...  
## $ real\_depart : POSIXlt, format: "2019-12-12 09:07:00" "2019-12-14 08:40:00" ...  
## $ expected\_arrival : POSIXlt, format: "2019-12-12 18:45:00" "2019-12-14 18:45:00" ...  
## $ real\_arrival : POSIXlt, format: "2019-12-12 19:34:00" "2019-12-14 18:45:00" ...  
## $ status\_depart : chr "Pontual" "Pontual" "Pontual" "Pontual" ...  
## $ status\_arrival : chr "Atraso 30-60" "Pontual" "Pontual" "Atraso 30-60" ...  
## $ observation : chr "CONEXÃO DE AERONAVE" NA NA "CONEXÃO DE AERONAVE" ...  
## $ delay\_depart : num 27 0 0 15 7 0 0 0 -1 0 ...  
## $ delay\_arrival : num 49 0 0 38 34 0 0 0 2 0 ...  
## $ expected\_flight\_length : num 605 605 605 605 605 605 605 605 605 615 ...  
## $ real\_flight\_length : num 627 605 605 628 632 605 605 605 608 615 ...  
## $ outlier\_depart\_delay : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ outlier\_arrival\_delay : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ outlier\_expected\_flight\_consistency: logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ outlier\_real\_flight\_consistency : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ outlier\_expected\_flight\_length : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ outlier\_real\_flight\_length : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ depart\_air\_temperature : num 21 23 26 21 24 20 23 23 20 23 ...  
## $ depart\_dew\_point : num 19 19 18 20 19 20 20 20 19 23 ...  
## $ depart\_relative\_humidity : num 88.3 78.2 61.4 94 73.6 ...  
## $ depart\_wind\_direction : num 270 90 20 130 320 320 70 90 60 100 ...  
## $ depart\_wind\_speed : num 2 5 9 5 13 2 6 3 9 3 ...  
## $ depart\_sky\_coverage : chr NA "SCT" NA "OVC" ...  
## $ depart\_pressure : num 29.9 30.2 30 30.1 30 ...  
## $ depart\_visibility : num 6.21 6.21 6.21 6.21 6.21 4.97 6.21 6.21 5.59 6.21 ...  
## $ depart\_apparent\_temperature : num 21 23 26 21 24 ...  
## $ depart\_wind\_speed\_scale : Ord.factor w/ 11 levels "Calm"<"Light Air"<..: 2 3 4 3 5 2 3 2 4 2 ...  
## $ depart\_wind\_direction\_cat : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 13 5 2 7 15 15 4 5 4 5 ...  
## $ arrival\_air\_temperature : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_dew\_point : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_relative\_humidity : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_wind\_direction : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_wind\_speed : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_sky\_coverage : chr NA NA NA NA ...  
## $ arrival\_pressure : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_visibility : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_apparent\_temperature : num NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_wind\_speed\_scale : Ord.factor w/ 11 levels "Calm"<"Light Air"<..: NA NA NA NA NA NA NA NA NA NA ...  
## $ arrival\_wind\_direction\_cat : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: NA NA NA NA NA NA NA NA NA NA ...

À primeira vista, observamos que o conjunto de dados possui diversos pontos que podem ser explorados. Para seguir um caminho de análise, vamos examinar algumas colunas específicas e, a partir delas, destacar certos aspectos de forma mais detalhada.

Por exemplo, as colunas expected\_depart, real\_depart, expected\_arrival e real\_arrival são variáveis temporais. Elas indicam os horários programados (expected\_) e os horários reais (real\_) de partida e chegada dos voos.

As informações de atraso ou adiantamento estão refletidas nas colunas delay\_depart e delay\_arrival. A coluna delay\_depart indica quanto tempo o voo partiu após ou antes do horário previsto. Quando o valor é negativo, como por exemplo -7, significa que o voo saiu 7 minutos adiantado. Já a coluna delay\_arrival mostra o tempo de atraso ou adiantamento na chegada, com base na diferença entre expected\_arrival e real\_arrival.

A lógica é que, quanto maior o atraso na saída, maior a probabilidade de ele impactar também o horário de chegada ao destino.

A coluna observation trás informações sobre o voo o que nos esclarece muitas coisas, sobre o que está acontecendo com o voo.

library(dplyr)

##   
## Anexando pacote: 'dplyr'

## Os seguintes objetos são mascarados por 'package:stats':  
##   
## filter, lag

## Os seguintes objetos são mascarados por 'package:base':  
##   
## intersect, setdiff, setequal, union

obs\_mais\_comuns <-dados %>%  
 filter(!is.na(observation)) %>%  
 count(observation, sort = TRUE)   
  
head(obs\_mais\_comuns, 20)

## observation  
## 1 LIBERAÇÃO SERV. TRAFEGO AÉREO/ANTECIPAÇÃO  
## 2 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 3 CONEXÃO DE AERONAVE  
## 4 ANTECIPAÇÃO DE HORÁRIO AUTORIZADA  
## 5 DEFEITOS DA AERONAVE  
## 6 CONEXÃO AERONAVE/VOLTA - VOO DE IDA NÃO PENALIZADO AEROPORTO INTERDITADO  
## 7 FALHA EQUIPO AUTOMOTIVO E DE ATENDIMENTO DE PAX  
## 8 CANCELAMENTO POR MOTIVOS TÉCNICOS - OPERACIONAIS  
## 9 CONEXÃO AERONAVE/VOLTA - VOO DE IDA NÃO PENALIZADO CONDIÇÕES METEOROLÓGICAS  
## 10 SEGURANÇA/PAX/CARGA/ALARME  
## 11 AEROPORTO COM RESTRIÇÕES OPERACIONAIS  
## 12 TROCA DE AERONAVE  
## 13 CANCELAMENTO - AEROPORTO DESTINO ABAIXO LIMITES  
## 14 FACILIDADES DO AEROPORTO - RESTRIÇÕES DE APOIO  
## 15 AUTORIZADO  
## 16 AEROPORTO ORIGEM ABAIXO DOS LIMITES  
## 17 VOO ESPECIAL DE RETORNO (EXCLUSIVO PARA RETORNO AO AEROPORTO DE ORIGEM)  
## 18 AEROPORTO DE ORIGEM INTERDITADO  
## 19 CANCELAMENTO - AEROPORTO ORIGEM ABAIXO LIMITES  
## 20 AEROPORTO DESTINO ABAIXO DOS LIMITES  
## n  
## 1 115375  
## 2 36107  
## 3 33877  
## 4 29174  
## 5 20834  
## 6 17274  
## 7 16282  
## 8 9311  
## 9 9169  
## 10 8692  
## 11 7652  
## 12 4378  
## 13 4196  
## 14 4158  
## 15 3457  
## 16 3031  
## 17 2598  
## 18 2174  
## 19 2166  
## 20 1998

Note que temos cerca de 45 tipos de observações diferentes, para reduzir a dimenção por problemas semelhantes vamos criar uma variavel com o proposito de agregar dados semelhantes

summary(dados$delay\_depart)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -525615.00 -5.00 0.00 -3.54 4.00 132895.00 18135

como podemos observar temos valores outliers

summary(dados$delay\_arrival)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -525623.0 -10.0 0.0 -4.3 2.0 305314.0 18135

vamos ver a distribuição geral

dados %>%  
 filter(delay\_depart >= -500, delay\_depart <= 500) %>%  
 summarise(  
 min = min(delay\_depart, na.rm = TRUE),  
 max = max(delay\_depart, na.rm = TRUE),  
 mean = mean(delay\_depart, na.rm = TRUE),  
 median = median(delay\_depart, na.rm = TRUE)  
 )

## min max mean median  
## 1 -499 500 4.789688 0

A mediana = 0 indica que mais da metade dos voos não tiveram atraso (ou partiram no horário). A média positiva (4.79 min) mostra que, no geral, os voos atrasam levemente. Os valores extremos originais (como 305314 ou -525623) provavelmente são erros ou registros inválidos.

library(ggplot2)  
library(dplyr)  
library(tidyr)  
  
tempo <- 100  
  
# Filtrar valores extremos para visualização  
dados\_filtrados <- dados %>%  
 filter(  
 delay\_depart >= - tempo, delay\_depart <= tempo,  
 delay\_arrival >= - tempo, delay\_arrival <= tempo  
 )  
  
# Transformar para formato longo  
dados\_long <- dados\_filtrados %>%  
 pivot\_longer(cols = c(delay\_depart, delay\_arrival),  
 names\_to = "tipo\_delay",  
 values\_to = "delay")  
  
# Plotar histograma com cores por tipo de atraso  
ggplot(dados\_long, aes(x = delay, fill = tipo\_delay)) +  
 geom\_histogram(alpha = 0.5, position = "identity", binwidth = 10, color = "white") +  
 scale\_fill\_manual(values = c("delay\_depart" = "steelblue", "delay\_arrival" = "darkred"),  
 labels = c("Partida", "Chegada")) +  
 labs(  
 title = "Distribuição dos Atrasos",  
 x = "Atraso (minutos)",  
 y = "Frequência",  
 fill = "Tipo de Atraso"  
 ) +  
 theme\_minimal()

library(stringr)  
  
dados <- dados %>%  
 mutate(obs\_categoria = case\_when(  
 str\_detect(observation, regex("LIBERAÇÃO|PLANO DE VOO|ANTECIPAÇÃO|AUTORIZAD[AO]", ignore\_case = TRUE)) ~ "Tráfego aéreo / Autorização",  
 str\_detect(observation, regex("DEFEITO|TROCA DE AERONAVE|AVARIA|PANE|DEGELO", ignore\_case = TRUE)) ~ "Problema na aeronave",  
 str\_detect(observation, regex("CONEXÃO AERONAVE|VOO DE IDA", ignore\_case = TRUE)) ~ "Conexão",  
 str\_detect(observation, regex("METEOROLÓGICAS|ABAIXO LIMITES|GELO|NEVE|LAMA", ignore\_case = TRUE)) ~ "Clima",  
 str\_detect(observation, regex("AEROPORTO .\*INTERDITADO|RESTRIÇÃO|FACILIDADES|DESTINO INTERDITADO|ALTERNATIVA", ignore\_case = TRUE)) ~ "Infraestrutura Aeroportuária",  
 str\_detect(observation, regex("EQUIPO|ABASTECIMENTO|DESTANQUEIO|OPERAÇÕES EM SOLO", ignore\_case = TRUE)) ~ "Falha de equipamentos/apoio",  
 str\_detect(observation, regex("SEGURANÇA|PAX|CARGA|ALFÂNDEGA|MIGRAÇÃO", ignore\_case = TRUE)) ~ "Segurança / Passageiros / Alfândega",  
 str\_detect(observation, regex("^CANCELAMENTO", ignore\_case = TRUE)) ~ "Cancelamento técnico ou climático",  
 str\_detect(observation, regex("FERIADO|VOO ESPECIAL|INCLUSÃO DE ETAPA", ignore\_case = TRUE)) ~ "Outros específicos",  
 str\_detect(observation, regex("ATRASOS NÃO ESPECÍFICOS", ignore\_case = TRUE)) ~ "Outros não específicos",  
 TRUE ~ "Outros não específicos"  
 ))  
  
atrasos\_por\_categoria <- dados %>%  
 group\_by(obs\_categoria) %>%  
 summarise(  
 media\_atraso\_depart = mean((delay\_arrival - delay\_depart), na.rm = TRUE),  
 total\_voos = n()  
 ) %>%  
 arrange(desc(media\_atraso\_depart))  
  
atrasos\_por\_categoria

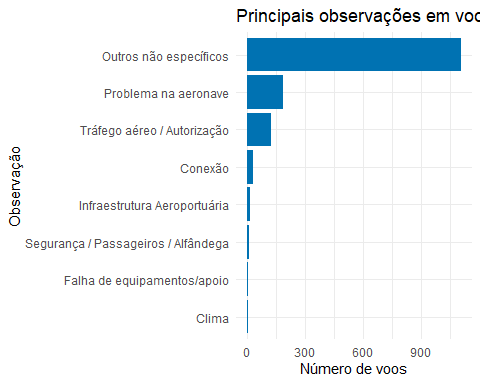
## # A tibble: 10 × 3  
## obs\_categoria media\_atraso\_depart total\_voos  
## <chr> <dbl> <int>  
## 1 Problema na aeronave 8.64 26026  
## 2 Clima 0.338 8352  
## 3 Outros específicos 0.00149 2729  
## 4 Outros não específicos -0.556 725134  
## 5 Conexão -0.771 27070  
## 6 Infraestrutura Aeroportuária -0.940 7248  
## 7 Falha de equipamentos/apoio -1.42 17889  
## 8 Tráfego aéreo / Autorização -3.11 149896  
## 9 Segurança / Passageiros / Alfândega -4.42 9741  
## 10 Cancelamento técnico ou climático NaN 9311

para enteder melhor o que está acontecendo com esses voo com o tempo de viagem tão longo, podemos observar seus atributos de tempo

library(dplyr)  
library(ggplot2)  
  
# 1. Filtrar voos com delay\_depart ou delay\_arrival > 500 minutos  
voos\_extremos <- dados %>%  
 filter(delay\_depart > 500 | delay\_arrival > 500)  
  
# 2. Ver as observações mais comuns  
voos\_extremos %>%  
 count(obs\_categoria, sort = TRUE)

## obs\_categoria n  
## 1 Outros não específicos 1111  
## 2 Problema na aeronave 186  
## 3 Tráfego aéreo / Autorização 124  
## 4 Conexão 32  
## 5 Infraestrutura Aeroportuária 15  
## 6 Segurança / Passageiros / Alfândega 12  
## 7 Falha de equipamentos/apoio 8  
## 8 Clima 7

voos\_extremos %>%  
 count(obs\_categoria, sort = TRUE) %>%  
 slice\_head(n = 15) %>%  
 ggplot(aes(x = reorder(obs\_categoria, n), y = n)) +  
 geom\_col(fill = "#0072B2") +  
 coord\_flip() +  
 labs(  
 title = "Principais observações em voos com atraso > 500 min",  
 x = "Observação",  
 y = "Número de voos"  
 ) +  
 theme\_minimal()



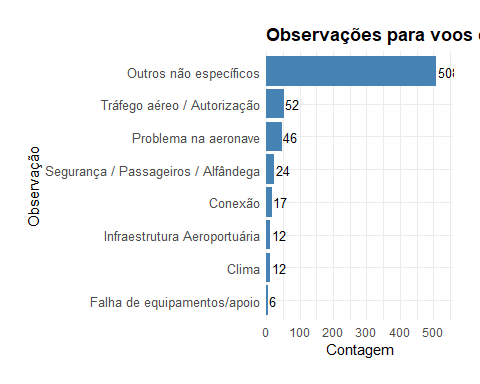
descrição de type: (N) Nacional, (I) Internacional, (R) Regional, (H) Sub-regional, (E) Especial, (c) frete/carga, (g) frete/carga internacional e (l) rede postal;

library(dplyr)  
  
voos\_atraso\_chegada\_grande <- dados %>%  
 filter(  
 delay\_arrival > 60,  
 !type %in% c("C", "G", "L"),  
 abs(expected\_flight\_length - real\_flight\_length) > 60  
 ) %>%  
 select(route, type, expected\_flight\_length, real\_flight\_length, observation)  
  
head(voos\_atraso\_chegada\_grande, 10)

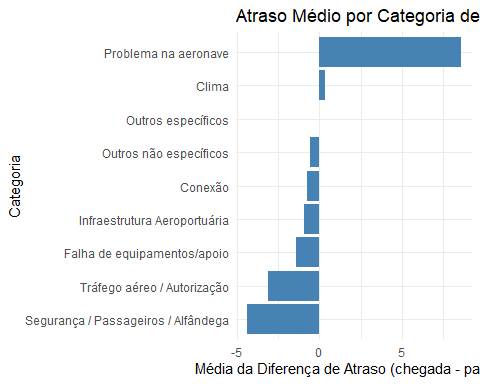
## route type expected\_flight\_length real\_flight\_length  
## 1 SBGR-CYYZ I 555 628  
## 2 SBGR-CYYZ I 615 546  
## 3 SBRF-EDDF I 600 904  
## 4 SBRF-EDDF I 600 922  
## 5 SBGL-EGLL I 680 1398  
## 6 SBGL-EHAM I 685 832  
## 7 SBGR-EHAM I 695 935  
## 8 SBGL-EHAM I 685 980  
## 9 SBGR-FAOR I 505 567  
## 10 SBGR-FAOR I 625 743  
## observation  
## 1 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 2 CONEXÃO DE AERONAVE  
## 3 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 4 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 5 CONEXÃO DE AERONAVE  
## 6 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 7 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 8 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 9 ATRASOS NÃO ESPECÍFICOS, OUTROS  
## 10 DEFEITOS DA AERONAVE

aqui podemos observar que mesmo com a demora para o voo sair e chegar a maior parte dos voos não tem tanta diferença de voo, porém se observamos os voos que tem uma janela de mais de 1 hora de expectativa de chegada temos o seguinte grafico

library(dplyr)  
library(ggplot2)  
  
obs\_counts <- dados %>%  
 filter(  
 delay\_arrival > 60,  
 !type %in% c("C", "G", "L"),  
 abs(expected\_flight\_length - real\_flight\_length) > 60  
 ) %>%  
 count(obs\_categoria, sort = TRUE) %>%  
 slice\_head(n = 10)  
  
ggplot(obs\_counts, aes(x = reorder(obs\_categoria, n), y = n)) +  
 geom\_col(fill = "steelblue") +  
 geom\_text(aes(label = n), hjust = -0.1, size = 3.5) +  
 coord\_flip() +  
 labs(  
 title = "Observações para voos com grande discrepância de tempo",  
 x = "Observação",  
 y = "Contagem"  
 ) +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(face = "bold"),  
 axis.text.y = element\_text(size = 10),  
 plot.margin = margin(t = 20, r = 20, b = 20, l = 20) # aumenta as margens  
 ) +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.1)))



library(ggplot2)  
library(dplyr)  
  
dados\_plot <- atrasos\_por\_categoria %>%  
 filter(!is.na(media\_atraso\_depart), !is.nan(media\_atraso\_depart), is.finite(media\_atraso\_depart))  
  
ggplot(dados\_plot, aes(x = reorder(obs\_categoria, media\_atraso\_depart), y = media\_atraso\_depart)) +  
 geom\_col(fill = "steelblue") +  
 coord\_flip() +  
 labs(x = "Categoria", y = "Média da Diferença de Atraso (chegada - partida)",  
 title = "Atraso Médio por Categoria de Observação") +  
 theme\_minimal()

 um detalhe muito interessante sobre as colunas status é que ela fala se o avião saiu no horário certo ou quando em qual ponto está o atraso do voo está em de 30 em 30 minutos

library(dplyr)  
  
dados\_contagem <- dados %>%  
 count(status\_depart, status\_arrival)  
  
dados\_contagem

## status\_depart status\_arrival n  
## 1 Antecipado Antecipado 371367  
## 2 Antecipado Atraso 120-240 70  
## 3 Antecipado Atraso 30-60 1070  
## 4 Antecipado Atraso 60-120 117  
## 5 Antecipado Atraso >240 64  
## 6 Antecipado Pontual 41140  
## 7 Atraso 120-240 Antecipado 32  
## 8 Atraso 120-240 Atraso 120-240 6186  
## 9 Atraso 120-240 Atraso 30-60 35  
## 10 Atraso 120-240 Atraso 60-120 821  
## 11 Atraso 120-240 Atraso >240 116  
## 12 Atraso 120-240 Pontual 41  
## 13 Atraso 30-60 Antecipado 434  
## 14 Atraso 30-60 Atraso 120-240 39  
## 15 Atraso 30-60 Atraso 30-60 27731  
## 16 Atraso 30-60 Atraso 60-120 2688  
## 17 Atraso 30-60 Atraso >240 20  
## 18 Atraso 30-60 Pontual 10697  
## 19 Atraso 60-120 Antecipado 51  
## 20 Atraso 60-120 Atraso 120-240 622  
## 21 Atraso 60-120 Atraso 30-60 2988  
## 22 Atraso 60-120 Atraso 60-120 14759  
## 23 Atraso 60-120 Atraso >240 18  
## 24 Atraso 60-120 Pontual 179  
## 25 Atraso >240 Antecipado 26  
## 26 Atraso >240 Atraso 120-240 186  
## 27 Atraso >240 Atraso 30-60 26  
## 28 Atraso >240 Atraso 60-120 30  
## 29 Atraso >240 Atraso >240 2974  
## 30 Atraso >240 Pontual 29  
## 31 Pontual Antecipado 93753  
## 32 Pontual Atraso 120-240 62  
## 33 Pontual Atraso 30-60 11277  
## 34 Pontual Atraso 60-120 481  
## 35 Pontual Atraso >240 65  
## 36 Pontual Pontual 375067  
## 37 <NA> <NA> 18135

para ficar melhor de vizualizar

library(ggplot2)  
  
niv <- c("Antecipado", "Pontual", "Atraso 30-60", "Atraso 60-120", "Atraso 120-240", "Atraso >240")  
  
dados\_contagem$status\_depart <- factor(dados\_contagem$status\_depart, levels = niv)  
dados\_contagem$status\_arrival <- factor(dados\_contagem$status\_arrival, levels = niv)  
  
ggplot(dados\_contagem, aes(x = status\_depart, y = status\_arrival)) +  
 geom\_point(aes(size = n, color = n), alpha = 0.7) +  
 geom\_text(aes(label = n), vjust = -0.8, size = 3.2) +  
 scale\_size\_continuous(range = c(2, 15)) +  
 scale\_color\_viridis\_c() +  
 theme\_minimal() +  
 labs(  
 x = "Status na Partida",  
 y = "Status na Chegada",  
 title = "Relação entre Status de Partida e Chegada",  
 size = "Contagem",  
 color = "Contagem"  
 ) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

